Title	物理学による脳と心の探求を始める好機到来 : 手作り国際研究集会"Toward a Science of Consciousness Tokyo '99"参加のお誘い&準備・運営への参画のお願い
Author(s)	保江, 邦夫
Citation	物性研究 (1998), 69(6): 826-850
Issue Date	1998-03-20
URL	http://hdl.handle.net/2433/96255
Right	
Type	Departmental Bulletin Paper
Textversion	publisher

物理学による脳と心の探求を始める好機到来*)

手作り国際研究集会
"Toward a Science of Consciousness Tokyo '99"
参加のお誘い
&
準備・運営への参画のお願い

ノートルダム清心女子大学
情報理学研究所
保江邦夫[1]
Ｔｏｋｙｏ’９９準備委員会[2]

（1998年2月17日受理）

１９９９年４月に東京にある国連大学で開催される予定の国際研究集会“Toward a Science of Consciousness Tokyo '99”へ向けて、若手の物理学者、特に物性物理学を専門とする皆さんに、この国際研究集会開催計画の学術的背景、準備状況、計画、展望などをご説明したいと思います。お読み下さった上で、もし物理学的な手法を前面に押し出した脳と心の探究に興味を持って頂けた場合には、是非にも会議に参加し、さらには会議の準備・運営にお力添えを頂戴できれば幸いです。

０．まずはじめに

物性研究の誌面をお借りして、私どもが計画している国際研究集会“Toward a Science of Consciousness Tokyo '99”（以後Ｔｏｋｙｏ’９９と略す）についてご紹介できるのは、編集委員の小嶋泉さんのご厚意によるものです。ここに、Ｔｏｋｙｏ’９９準備委員会を代表し

---

[1] 連絡先：E-mail kunio@ndsu.ac.jp　Tel. 086-252-1155（Ext.351）　Fax.086-255-5090
[2] メーリングリスト：　cons99@zoushoku.narc.affrc.go.jp　（English text only）

*）本稿は、編集部の方から特にお願いして執筆していただいた記事である。

て御礼申し上げます。

まずは、この辺りのことから説明していきたいと思います。ほとんどの方がご存じのように、小嶋さんはゲージ場の相対論的な正準理論の構成に始まり、場の量子論や多体系の量子力学（統計力学を含む）についての数理物理学的な研究をリードしてこられた理論物理学者です。彼の学問的興味には幅広く、かつ奥深いものがあり、その片鱗は物性研究の編集後記にすら見え隠れしています。何年か前まで基研や数理解析研で開かれていた「進化の力学についての諸問題」と題する研究集会を組織し、従来の物理学の縄張りにとらわれることのない新しい数理科学の勃興を促す原動力となったことも、ひとえに小嶋さんの旺盛な探求心とオープンマインディッドなお人柄のたまものでしょう。

私も、研究集会「進化の力学についての諸問題」には何度か出席しましたが、一度だけ脳における記憶の素過程を量子場の自発的対称性の破れの現象として捉える話をさせていただきました。もう７、８年前のことなので完全な記憶からは程遠いのですが、いわゆるニューラルネットが大流行の時期で、会場にも神経回路網の古典論をやっている人が多くいたと思います。当然のように、その人たちからは「脳の素過程の研究に量子論の必要なし！」というささか感情的なご意見を頂戴したのですが、私自身はそのようなご意見よりも果たして小嶋さんがどうお考えになるのかだけが気になっていて、いちいち反論することもしませんでした。（京都の旧友に言わせると、ちょうど「牙の抜けた保江君」状態になっていたとはいえ、いまから思えば「この自然界の全ては量子論で記述され、古典論で記述されるのは近似的に見た場合のみである」ことくらいは進言していたほうがよかったかもしれません）

ともかく、なぜ小嶋さんの反応だけに全神経を注いでいたかともうしますと、むろん場の量子論や量子統計力学を熟知していらっしゃるだけに、脳の素過程としての記憶のメカニズムを場の量子論の自発的対称性の破れによって記述しようとする私どもの発想の是非を即答されてしまうのではないかという心配もありましたが、それよりも場の量子論の専門家である前に医学者でもあった彼の医学的センスにより一刀両断されてしまうのではないかという不安が先立っていたからなのです。幸いにも（？）彼のコメントは、実際の脳は体温下の熱雑音を受ける物理系として考えなければならないから、有限温度の場の理論で記述する必要があるのではないかという建設的な（おだやかな？）ものでした。

私はその場で即答できなかったのですが、その後５、６年をかけてリファインした私どもの理論では、脳の素過程を担っている量子場は頭蓋内の電磁場であり、脳細胞組織や近傍の秩序化された運動状態にある水分子が持つ電気双極子の平均場が自発的対称性の破れを持つことから、ヒッグス機構をとおして１２～１３ｅＶ程度の有限質量を持つ光子（エバネッセント光子）が巨視的凝集状態を実現していることまで突き止められました。このエバネッセント光子のボーズ凝縮状態の臨界温度を理想ボーズ気体として近似計算すると、摂氏で３８～４０度程度となり、体温下の脳組織においても巨視的量子効果の影響が本質的であること

が分かります。

というわけで、場の量子論に慣れ親しんでいる理論物理学者の前でならば、ようやく胸を張って「脳の素過程は巨視的量子効果だ！」と主張できるだけの素地ができあがったのですが、残念ながら物理学者の中にも必ずしも場の量子論に強い人たちだけがいるわけでもありませんし、ましてや脳組織の複雑な物理現象を複雑系として古典近似の中で対症療法的に調べていくことが物理的なアプローチだと考えている情報科学や認知科学の人たちには、量子力学も量子論も場の量子論も全ては同じ穴の狢としか映らないのです。ですから、脳科学や認知科学の専門家に私どもの研究の内容を伝えるときには、それこそ最新の注意を払う必要があるわけです。

1．混沌からの出発

私どもは、脳細胞組織の中にエバネッセント光子のボーズ凝縮状態が実現されていることを理論的に示すことができたのですが、どうも世の中の巡り合わせがよくなかったようです。運の悪いことに、英国の女性でダナ・ゾーハという（自称）物理学者が「人間の自我は脳の中に実現したボーズ・アインシュタイン凝縮により発生している」という持説を「クォンタム・セルフ（Quantum Self）」という通俗本の中で取り上げたため、ボーズ・アインシュタイン凝縮が脳組織の中にあるというだけでこのゾーハと同じデタラメな説だと信じ込まれてしまうのです。もちろん、ゾーハの説の中で彼女が勝手に思い描いている脳の中のボーズ・アインシュタイン凝縮とは、フレーリッヒの生体フォノンや生体高分子中の電子とホールのペアーでできたボーズ粒子についてのものでした。そのような物質場のボーズ・アインシュタイン凝縮体が実現したとしても、その臨界温度は極低温となり、そもそも生体細胞組織自体が生きて存在し得ない環境となっているため、ゾーハの通俗的な説に対して多くの健全な物理学者達が反論していたのも納得できます。

しかし、物理学のことを知らない脳科学や認知科学の専門家にとっては、一度ほとんどの物理学者が否定したゾーハの脳の中のボーズ・アインシュタイン凝縮も、新たに登場した私どもの脳の中のエバネッセント光子の凝集体も同じものとしか映らず、結果として変な目で見られてしまいかねません。

さらに都合の悪いことに、何年か前にやはり英国の鬼才で一般相対論や宇宙論の研究にトポロジーの手法を持ち込んで意気盛んなロジャー・ペンローズが、これまでの人工知能研究の延長線上には人間の意識についての真の理解はあり得ないという警鐘を鳴らした成書を立て続けに２冊も出版したため、意識の研究についての世界的な論争に火がついてしまったのです。ペンローズは、人間の意識の根底にある素過程はチューリングマシンでは絶対に模倣ができない非計算的アルゴリズムを実現しているが、これまで知られている物理的素過程の

中でこのような非計算的要素を持つものは量子力学における観測に伴う量子状態の収縮過程しかないということで、人間の意識のメカニズムの基本は何らかの量子力学的な体系によって与えられているはずだと考えたのです。

量子力学の観測問題自身、まさに人間の自由意志の問題と絡められ、フォン・ノイマンやウィグナーなどの天才的数学者や物理学者等によって観測に伴う量子状態の収縮には観測者の意識が明示的に作用していると議論されたこともあったため、これがペンローズの主張についての受けとめる側の混乱をいっそう助長してしまったことも事実です。

というわけで、誰も何も正しく理解せぬ間に、ペンローズの主張のうち、脳の中には量子力学系があって、観測によるその量子状態の収縮過程により意識の素過程が生じるというものが一人歩きし始め、いまやちゃんとした物理学者が何を言っても収拾がつかないほどに混乱した議論がインターネットを飛び交い、もはや手の施しようがありません。

一時は、世界的に広まったこのような無益な論争から離れ、自分達だけはちゃんとした研究を続けようとインターネットでの議論から手を引いてはみたのですが、私どもの目が行き届かぬうちにペンローズまでもが混乱し始め、孤立していない限りどんな系も量子力学では記述できないという言葉どおりの解釈を前面に押し出し、細胞の中に孤立した量子力学系を探し求め始めるありさま。このままでは、本当のことは埋もれてしまい、結局は声の大きかった数人の人たちの意見が正論と見なされてしまいかねません。

一昨年くらいから、このような状況を憂う何人かの若者達が私どもの所に押し掛けてくれ始め、一度きちんとした物理学の立場から脳だけでなく生体系の中での巨視的量子効果として理解できる素過程を明示した上で、人間の意識の本質的理解に果たして本当に量子物理学の枠組みが必要か否かを真剣に討議し、将来のこの方面での研究の確固たる足場となるような国際会議を東京で開くべきだという話題が出始めていました。

２．手作り国際会議？

といっても、大きな学会や研究団体のバックがあるわけではない私どもができることといえば、たかが知れているのも事実です。しかし、何もせずに手をこまねいているだけでは、事態は悪化する一方ではないかという声に押し切られたという形で、昨年の１１月にまだ先が見えない状態のまま取りあえずＴｏｋｙｏ’９９開催準備委員会を発足させました。というわけで、準備委員会ができてからの経緯の概略、特に国連大学との関わりについては後程触れることにしますが、ここではまず、私自身がこの国際会議開催に向けた決断をすることができた経緯をご紹介したいと思います。

ご存じの方も多いと思いますが、今年の１月４日から８日までの日程でフィリピンのボホール島にあるハグナ市において "The Second Jagna Workshop on Mathematical Methods of

Quantum Physics" と題する国際会議がありました。江沢洋先生の６５才のお祝いを兼ねていたため、日本からも１０名以上の参加がありましたし、講演内容も多岐にわたる興味深いものがあり、鄙びた南の島で夜更けまで議論が沸きました。久しぶりにお会いした小嶋さんも最新の結果を講演され、その後の食事のときなどに色々とお話をうかがうことができ、たまたま私どもが１９９９年に東京で国際会議を開こうと考えているということまで話題に登りました。

会期中の一晩は "Ezawa Festival" ということで、江沢先生の６５才の誕生日をお祝いする宴が催され、アメリカのクラウダーやドイツとポルトガルのシュトライトが気の利いた祝辞を披露した後、ひょんなことから私にマイクが回ってきました。冷や汗を流しながら、とにかくお役目を果たした私は少し落ち着きを取り戻し、最後にこのボホール島での国際会議の運営のすばらしさに感銘したことについて、正直な気持ちをご披露できたのです。実は、この国際会議の計画から準備、運営、それに食事や宿泊を含めた裏方的な作業のほとんどを、主催者のベルニード教授の家族や縁者が分担して行っていました。いわば手作り国際会議といえるような、質素でそれでいて開催者の暖かく細かい配慮が行き届いた、居心地のよい環境の中での５日間は、私どもが忘れかけていた原点を思い出させてくれたのです。

Ｔｏｋｙｏ’９９の準備計画としては、それまでは東京で開催する国際会議ということで、産・官・学の主だった組織や団体を背景として、大手の情報通信関連の企業やマスコミの援助を得た上でそれなりの立派なものにしなければいけないと思い込み、私ども準備委員会は各方面への働きかけを始めていました。しかし、ボホール島での手作り国際会議を目の当たりにして、Ｔｏｋｙｏ’９９の理想像は東京で開催するからといってそのような上辺の華やかさや形に求められるべきものではなく、むしろ東京でもその気になれば手作りに近い形の質素で居心地のよい国際会議を成功させることができることを実証することにあるという確信を得たのです。

３．国連大学の協力

質素でなおかつ暖かく細かい配慮が行き届いた、居心地のよい国際会議を東京で開催するといっても、とても私どもの力だけでは無理なことは明かです。一番楽なやり方は、最近日本でもこの方式の国際会議が目立ってきていますが、企画から運営にいたる全てを請け負ってくれるコンファレンス・コーディネーターの会社に依頼してしまうことでしょう。しかし、これではせっかくのボホール島での啓示も、水泡に帰すこととなります。あくまで、私どもの力の及ぶ範囲でだけで手作りの国際会議に仕立て上げたいわけですから。

となると、会議の運営自体も純粋に参加者の負担だけに頼らざるを得なくなりますが、それでは大規模な国際会議が開けるような立派な会議施設を東京の都心で５日間にわたって使

用するための費用を捻出することすら無理になってしまいます。東京にあるどこかの大学の施設をお借りすることも考えましたが、なかなか大規模な国際会議が開けるような場所や、その運営のための便宜を提供していただくことには無理があるようでした。準備委員会のメンバーで長時間議論していたのですが、アイデアも出尽くし、いささか困り果てて小休止をとっていたところ、メンバーの中でも若くて元気な治部眞里さん（ノートルダム清心女子大学情報理学研究所）と平藤雅之さん（農林水産省研究情報部モデル開発研究室）がインターネットのウェブの中から国連大学のサイトを捜し当て、「これだ！」と盛り上がってきたのです。

見ると、東京の青山の超一等地に、立派なビルと大規模な国際会議を開くための施設が整っているではありませんか。「国連大学といっしょにやりさえすれば、この素晴らしい施設を使わしてもらえるはずだ！」そう考えた私どもは、さっそく次の週にはメンバーの中の唯一の外国人であるヘイガンさんを前面に押し出し、背水の陣で国連大学にお願いに行ったのです。まさに、窮すれば通ず。私どもの熱意が国連大学高等研究所のデラ・センタ所長に伝わり、氏のご厚意でＴｏｋｙｏ’９９を国連大学で開催させていただけることになったのです。

これで、Ｔｏｋｙｏ’９９に国内外から３００人程度の方々が参加して下されば、何とか参加費だけでこの手作り国際会議を東京で開催することができる目途が立ちました。残るは、果たしてそのように大勢の方々に興味を持っていただけ、参加していただけるかという危惧だけになったのですが、これについてはＴｏｋｙｏ’９９の主旨や開催の背景、計画の詳細な内容などについて、できるだけ多くの方々にご説明する機会を持つしかありません。今回、編集委員のお一人である小嶋さんのご厚意で「物性研究」の誌面をお借りして、日本の若い世代の物理学者の皆さんにご説明させていただけることになりました。皆さんの中から一人でも多くの方が「脳と心の物理学」に興味を持って下さり、Ｔｏｋｙｏ’９９に参加、あるいはお手伝いいただけるならば、準備委員会の一員としてこれに勝る喜びはありません。

どうか、よろしくお願いいたします。

Ｔｏｋｙｏ’９９の詳細な開催計画は本年３月末日に最終決定する予定ですので、ここでは残念ながら８０％程度の暫定的な開催計画しかご紹介できません。しかし、ここでご紹介する内容と大きく違ってくるような場合には、改めて本誌をお借りしてご報告する予定ですので、悪しからずご了承下さい。

なお、Ｔｏｋｙｏ’９９開催の学術的背景の解説につきましては、準備委員会のメンバーの一人であるヘイガンさんが過日国連大学高等研究所で行った講演内容を基に解説論文を書いてくれましたので、最後にそれを掲載させていただくことで代えさせていただきます。また、“脳と心の物理学”についての最新の話題や将来の展望について、特に巨視的量子効果の重要性を主張したものとして

“脳と心のバイオフィジックス”（日本生物物理学会編、共立出版、１９９７）収録、保江邦夫担当執筆「心の量子論」

“Quantum Brain Dynamics and Consciousness – An Introduction” (John Benjamins, Amsterdam, 1995) ISBN 1-55619-183-9

“１リットルの宇宙論－量子脳力学への誘い”（治部眞里・保江邦夫共著、海鳴社、１９９１）

などの成書がありますので、ご覧ください。

また、量子力学と時空構造の統合の中に心の自由度が入るべきだという唯心論的な主張を展開したものとして

“唯心論的物理学の誕生－モナド・量子力学・相対性理論の統一モデルと観測問題の解決”（中込照明著、海鳴社、１９９８）

も参考になると思います。

４．Ｔｏｋｙｏ’９９準備計画

準備委員会（１９９７年１１月６日発足）：

Stuart Hameroff (Arizona, Medical Science)
Roger Penrose (Oxford, Mathematics)
Tarcisio Della Senta (Tokyo, UNU/IAS)
Naoyuki Osaka 苧阪直行 (Kyoto, Cognitive Science)
Masayuki Hirafuji 平藤雅之 (Tsukuba, Interdisciplinary Science)
Scott Hagan (Tsukuba, Physics)
Ng S. T. Chong (Tokyo, UNU/IAS)
Mari Jibu 治部眞里 (Okayama, Brain Science)
Kunio Yasue 保江邦夫 (Okayama, Physics)

会議名称：

Toward a Science of Consciousness - Tokyo '99
(Abbreviation: Tokyo'99)

会議主題：

What is ultimately responsible for the phenomenon of consciousness?
Philosophical, neurobiological, cognitive, physical, and computational views.

主催：

Tokyo'99 Organizing Committee
The Institute of Advanced Study of the United Nations University
The University of Arizona Center for Consciousness Studies

開催期間（予定）：

From 19 (Monday) to 23 (Friday) of April, 1999

開催場所：

The United Nations University (Tokyo/Aoyama)

会議形式：

Plenary Session + Poster Session + Symposium
The 5th Day (Friday, April 23) will be devoted to a Symposium with 10 minute talks by all the keynote speakers and coordinators.
Plenary SessionとSymposiumは収容人員３００人のホールで行い、Poster Sessionは小ホールで行う。

開催目的と展望：

"Toward a Science of Consciousness - Tokyo'99" is an international, interdisciplinary forum for the application of scientific concepts to brain functioning at all levels and to consciousness. Keynote talks by the world-wide most active researchers in the fields of brain science, cognitive science, physical science, mathematical science and philosophical science are given, followed by the detailed presentations of the newest results of selected original experimental and theoretical investigations.
The purpose of the forum is not only to provide the wider audience with the correct image of the actual research activity from the scientific point of view aimed at revealing WHAT CONSCIOUSNESS IS, but also to realize good and meaningful encounters among different scientific researchers from different research fields and countries so that the seemingly delicate findings from the future CONSCIOUSNESS SCIENCE would be made open to and shared by all the people living on this planet.
The forum is, therefore, also dedicated to the extensive discussion and preparation

of the 1999 Tokyo Declaration by all the participants that CONSCIOUSNESS SCIENCE WILL BE ONLY FOR HUMAN WELFARE BUT NEVER FOR WARFARE.

準備委員会連絡先：

Tokyo'99 Organizing Committee
c/o
Dr. Masayuki Hirafuji
Head, Computational Modeling Laboratory
Department of Information Science
National Agriculture Research Center
3-1-1 Kan-non Dai, Tsukuba, 305 Japan
E-mail hirafuji@narc.affrc.go.jp
Tel. 0298-38-8986
Fax. 0298-38-8551

会議録：

The Conference Proceedings containing all the submitted papers from the participants will be published in both printed and electronic forms.

主たる講演予定者：

本年4月1日までに最終決定を行う予定で現在交渉中ですが、既に内諾を得た外国からのものについて列挙しておきます。

Roger Penrose（数学、意識の科学的研究と量子効果の関連についての火付け役）
Stuart Hameroff（医学、麻酔を専門としながら、ペンローズとともに意識の科学の研究を始めた）
Nicolas Gisin（物理、量子暗号、量子コンピューティング、観測問題）
Gordon Globus（精神医学、哲学）
Karl Pribram（大脳生理学、記憶のホログラフィー理論）
Giuseppe Vitiello（物理、場の量子論）

これ以外には、次のような分野の専門家による講演を予定しています。

量子コンピューティング

量子ニューロコンピューティング
ニューラルネットワークモデルとしての力学系
人工知能
人工生命
認知科学
生物物理学
量子生物学
科学・医学倫理

５．解説論文

Ｔｏｋｙｏ’９９開催の学術的背景の解説論文として、準備委員会のメンバーの一人であるヘイガンさんが本年２月５日に国連大学高等研究所で行った講演の内容を基にしたものを以下に掲載させていただきます。

Physical and mathematical theory in a scientific approach to consciousness

S. Hagan
Computational Modeling Lab,
Dept. of Information Research, N.A.R.C.
3 - 1 - 1 Kannondai, Tsukuba, Ibaraki, 305 Japan

1 Introduction

Numerous disciplines have, in the past decade, begun to lay the foundations of a multidisciplinary synthesis, forging the basis for a scientific understanding of the origins of consciousness. Until recently physicists and mathematicians have remained conspicuously absent from this effort, no doubt at least in part because of the belief, widely held both amongst the general public and in academic circles, that functional accounts of consciousness at the neural or cognitive level should prove entirely adequate to an explanation of the phenomenon. That perception has now been questioned by the emergence of several converging lines of argumentation, proceeding from fundamental logical and physical constraints, that challenge the most basic assumptions of our current approach to understanding how the brain gives rise to the mind.

2 Consciousness as a Natural Phenomenon

For a phenomenon whose existence is unlikely to be sincerely questioned by anyone, 'consciousness' has, until recently, received scant serious attention from the scientific community. The view is commonplace in academia that proper scientific detachment requires that the legitimacy of a scientific research project to study consciousness be deemed highly suspect. Science, it is held, must study only the facts and subjective experiences are not the stuff of facts. Facts are objective things and are open to third-person confirmation. Consciousness is neither.

While this position would seem to make the case in a rather straightforward and unobjectionable manner, the philosophical foundation on which it rests collapsed utterly several decades ago. One would be hard-pressed to find even a single adherent, amongst contemporary working philosophers, of the philosophical position, known as logical positivism, that it represents. Logical positivism flourished in the early part of this century and won many advocates from the scientific disciplines by contending that any statement about matters of fact not verifiable, at least in principle, on performing some conceivable experiment, was a meaningless statement. It is easy to see how this might be applied to the mind and how it might relegate consciousness to the metaphysical wastebasket[17].

This line of reasoning had a rather compelling impact on scientific history by persuading psychologists of the behaviorist school to focus their efforts, not on the mind *perse*, but on the patterns of stimulus and response that supposedly led us to postulate a mind in the first place. For the span of several decades the 'science of the mind' lost sight of its intended object and eventually proscribed all mention of what might take place inside the black box that was the brain. The prohibition on consciousness has in some respects persisted to this day; the difficulty we experience uttering words like 'mind' or 'consciousness' in the company of scientific peers is a legacy of this era. These words connote no mysticism nor conjure any supernatural beings but merely acknowledge a fact, a fact recognized since Descartes as beyond reasonable doubt.

The demise of logical positivism proceeded from this very fact. If the existence of conscious states is as indubitable as Descartes claimed, then according to the logical positivists there ought to be some imaginable experiment in which they could be detected. The meant distinguishing the conditions of satisfaction for a particular conscious state to obtain in terms of an unequivocal and objective behavioral signature. Yet it was always logically possible to imagine, at least in principle, a superlative actor whose performance would give away nothing of his actual thoughts, whose outward behavior would disguise rather than reveal the underlying conscious state[16]. This meant that behavior was insufficient to determine consciousness, that there really were facts that were not open to third-person confirmation.

These developments in philosophy seemed to have escaped the notice of those within the artificial intelligence circles advocating the use of the Turing test. This test was once, and sometimes still is, considered the ultimate guide to success or failure in the project of creating a computer capable of thought.

The test is simple: a candidate computer is deemed conscious if its responses to questioning cannot be distinguished from a person's responses. The persistence of the Turing test as a philosophical anachronism is at least in part due to an ambiguity in terms[18]. The word 'intelligence' that occurs in 'artificial intelligence' is sometimes discussed as if it implied the same thing as consciousness, replete with thoughts, perhaps even mental imagery. In other contexts the word connotes nothing more than 'goal-directed behavior', a quality that one might easily grant to a thermostat without implying that it was "thinking about the temperature." If 'goal-directed behavior' is intended however the need for a Turing test never arises, since there is no implication of a thinking existence underlying the behavior. If actual consciousness is intended, the test is ill-conceived, since consciousness is not susceptible to third-person observation.

But is consciousness then a suitable object for scientific research? If it is not, even in principle, possible to make third-person observations of the phenomenon, how can it be admitted as an appropriate research topic? Does this not contradict some central dogma of science? If the view is taken seriously that science must admit only such facts as are open to direct, third-person confirmation, it would lead to the utter demise of accepted fields of study like cosmology or evolutionary biology. The objects of these sciences are veiled in prehistory. The beginning of the universe and the origins of species, like consciousness, are black boxes to which we presumably have no direct experimental access. How does research proceed in these areas? One considers an over-arching theory and examines the ramifications produced within the realm of observable phenomena. The theory is updated whenever new empirical data becomes available but is equally conditioned by information concerning the internal consistency of its presuppositions or the validity of the phenomenological procedures for determining the empirical consequences of the theoretical processes and entities. This gives a hint as to the direction that might be pursued by a scientific study of consciousness. It suggests a research strategy which proposes a theoretical basis of explanation and conditions the theory on its consistency with the features of the phenomenon as well as with empirical observation of the indirect effects.

3 Considerations in Mathematics

Amongst those researchers willing to grant the legitimacy of a scientific research project to study consciousness, there are those who claim that such a project is already well underway, carried out under the auspices of neurophysiology and cognitive science; that a neural or functional account of the activities of the brain should prove entirely adequate to an explanation of consciousness; that no additional considerations, beyond the purview of these disciplines, need arise when one expands the scope of investigation from the purely cognitive features of the brain's functionality to the phenomenon of consciousness itself. On this view there is in fact no distinction between the physical brain seen in its computational or functional capacity and the mental features we associate to consciousness.

The cracks in the computationalist thesis have been evident for some time[19] but have recently become catastrophic with the introduction of an argument, championed by Penrose[15], that claims to provide a mathematical demonstration of the insufficiency of *any* purely computational explanation of the inner workings of 'human understanding'. In light of the potential import of such a result to a scientific conception of what the brain does, the basis of the argument is reviewed here.

In order to formulate a mathematical refutation of the computational nature of human understanding, it might seem that Penrose must first rigorously define the term 'human understanding', an exercise sure to excite considerable controversy. Rather than entering such a debate, Penrose devises his argument to prove only that one particular human activity is not susceptible to algorithmic formulation. A single counterargument is sufficient for Penrose to make his case for if even one of the activities in which humans engage can be proven to lie outside the bounds of a computational account, the entire computationalist project is derailed.

The argument then focuses on the particular faculty of understanding that humans engage to determine *whether or not a given program, once implemented, will ever reach completion* - a question reknowned in computer and cognitive science as the *halting problem.* The problem harbours some grave implications in the foundations of mathematics, implications of which the mathematical community became aware in 1931 following the release of a paper written by Gödel[4], a paper that proved to be a Trojan Horse for the formalist school of mathematics that solicited it. Gödel's Incompleteness Theorem dealt a decisive blow to the formalist school by exposing in its premises a fatal flaw, a defect that could not be rectified from within. Penrose's argument, crafted in the same style of reasoning, seeks to expose a similar error in the fundamental assumptions of the computationalist school, an error that urges the abandonment of the computationalist project.

As in Gödel's original work, the argument begins by specifying a particular computation to be assessed. For our purposes, it will be sufficient to consider only those computations which take a single input. All the possible computations that can be performed on a single input variable can be numbered sequentially in an unambiguous fashion and, as long as the numbering scheme adopted is used consistently, the particular order in which the computations are arranged is of no consequence. Under some such scheme, the qth computation with an input variable n can be denoted:

$$C_q(n) \qquad \text{for } q, n \in N. \tag{1}$$

If it is possible to computationally represent all the various aspects of human understanding brought to bear by the research community in determining whether or not this particular computation will ever reach completion, then this understanding can be represented algorithmically. The algorithm:

$$A(q, n) \qquad \text{for } q, n \in N, \tag{2}$$

then encapsulates, by hypothesis, all human means to decide if the qth computation acting on input n will halt. Such an algorithm will halt only when it can be computationally ascertained that the given computation with the specified input does not come to a halt:

$$A(q, n) \text{ halts } \Rightarrow C_q(n) \text{ does not halt.} \tag{3}$$

The direction of the implication should be noted from the outset. That is, if $A(q, n)$ stops, it is certain that the computation $C_q(n)$ will not stop but if $A(q, n)$ does not stop, no information whatsoever is communicated regarding whether or not the computation $C_q(n)$ stops.

Since the input n is a free variable, it is possible to assign it any value. In particular one could assign as input to the qth computation the very same value q. With this assignment, the computation becomes $C_q(q)$ and the algorithm, encapsulating all human means of determining if that computation halts, is $A(q, q)$:

$$A(q, q) \text{ halts } \Rightarrow C_q(q) \text{ does not halt} \tag{4}$$

When the two arguments of the algorithm A are given the same value in this way, the algorithm is reduced to a single input. As such, one must now expect that the algorithm, in the particular case for which the computation number and the input number are the same, is itself identical to one of the computations in the ordered list of *all* computations on a single input. This list, after all, is exhaustive. Let us say that the algorithm, when reduced to a single argument, is found to be identical to the kth computation:

$$A(q, q) = C_k(q) \text{ for some particular value of } k \tag{5}$$

$C_k(q)$	assesses	$C_q(q)$
$C_k(1)$	assesses	$C_1(1)$
$C_k(2)$	assesses	$C_2(2)$
$\vdots$	$\vdots$	$\vdots$
$C_k(k)$	assesses	$C_k(k)$

Table 1: C_k is *always* the assessing algorithm when the computation number and input number of the assessed computation are identical.

Note that C_k is always identical to the algorithm whenever the computation number and input number of the program to be assessed coincide. The input variable q given to C_k specifies both the computation number *and* the input number of the computation to be assessed.

Since the input q is still a free variable, it is possible to assign it any value. In particular, we might assign it the value k found above:

$$A(k,k) = C_k(k) \text{ for some particular value of } k, \tag{6}$$

and, substituting k for q, the implication in (4) becomes:

$$A(k,k) \text{ halts } \Rightarrow C_k(k) \text{ does not halt.} \tag{7}$$

A glance at Table 1 indicates the significance of this value for q. The algorithm has been assigned to assess itself.

Using identity (6) to substitute for $A(k,k)$ in implication (7), one obtains:

$$C_k(k) \text{ halts } \Rightarrow C_k(k) \text{ does not halt,} \tag{8}$$

We can now make some logical observations that will allow us to decide the question of whether or not this computation, the algorithm itself, will halt. If $C_k(k)$ *does* halt, then the last statement above indicates that it must also *not* halt - a logical impossibility. If $C_k(k)$ does *not* halt, no contradiction is implied so we must conclude that $C_k(k)$ does not halt *reductio ad absurdum*:

$$C_k(k) \text{ does not halt} \tag{9}$$

Armed with the knowledge that $C_k(k)$ does not halt, we now examine how the algorithm fared in determining whether or not this program would halt. This algorithm after all encapsulates by hypothesis all the powers of human reasoning that we might bring to bear on this problem. If it is possible to give those powers a computational form, then we should expect the algorithm to come to the same conclusion that we did.

The algorithm to assess the program $C_k(k)$ *is* the very same program $C_k(k)$, a computation that we know does not halt. The *algorithm* does not halt. It is therefore never able to return an answer and the question whether or not $C_k(k)$ halts remains *computationally* undecided. Since we nevertheless *know* that $C_k(k)$ does not halt, it is evident that we did not arrive at this knowledge by purely computational means.

The power and scope of this kind of reasoning should not be underestimated. It is not for instance undermined by multiple, dynamic or 'bottom-up' computational strategies. Multiple algorithms can *always* be woven into a single algorithm that halts only when *all* of the component algorithms halt, and this new algorithm will once again be susceptible to the argument. Dynamic programs must follow a set of procedures which determine the course of change. If *these* are entirely algorithmic they must naturally be included in the specification of the algorithm A and the argument proceeds as before. Finally the characterization of computation in the argument makes no distinction between 'top-down' and 'bottom-up' processing. These are design tactics, descriptions of the method by which one chooses a particular computation, but are utterly irrelevant to the mere listing of all possible computations.

Perhaps the most common misconception of the Gödelian style of reasoning proceeds from the observation that the procedure by which we obtained the insight that "$C_k(k)$ does not halt" appears entirely algorithmic and it should therefore be possible to give it a computational form. This assertion is in fact correct. There is of course a computational procedure to obtain the particular computation $C_k(k)$ given the algorithm A. One might imagine that such a procedure would allow the algorithm A to find the very computation that Penrose's argument singles out. The program could then be instructed that this particular computation is non-halting. The implications of Penrose's argument for the computationalist camp are not so easily averted however. This strategy fails because the procedure to find $C_k(k)$ given A is not a procedure in A - it *cannot* be since A, as we have seen, is incapable of deciding whether $C_k(k)$ will halt whereas an algorithm newly equipped with a procedure to find $C_k(k)$ *is* capable of deciding the issue. One might then suggest that the procedure be appended to the old algorithm A to form a new algorithm A' capable of finding $C_k(k)$. But the particular computation $C_k(k)$ which Penrose's argument specifies for this new algorithm A' will not be the same as the one for the old algorithm A. The new algorithm will be undecided about a different computation and will, in its turn, be unable to locate that computation.

The realm of pure mathematics may seem a strange place to turn for answers regarding the nature and origins of consciousness, a phenomenon firmly rooted, at least to this point in history, in complex biological processes. But computer science might once have seemed to have little to offer a mature science of neurophysiology towards the goal of elucidating those same biological processes. Such offerings are of course now commonplace. Natural phenomena are not required to respect the artificial borders prescribed for science. Researchers, like Penrose, currently contributing their theses to a multi-disciplinary understanding of the scientific basis of consciousness do not argue that neurophysiology and cognitive science have nothing to do with generating conscious states. Such a proposition is clearly ludicrous in the face of overwhelming evidence that neural activities in the brain are closely correlated with mental phenomena. But the obvious *necessity* of the understanding provided by these particular disciplines is not proof of their *sufficiency*. And it is on this point, and this point alone, that they are questioned. Penrose, for instance, does not make the claim that consciousness comes about or even could come about in the absence of calculation. He is rather pointing to the insufficiency of computational accounts and insisting that some room must be made somewhere for a non-computational component.

Much of the reaction to the thesis has focused, not on the finer points of the argument itself, but on the inferences that Penrose draws from its conclusion and the direction he indicates for further research. Contemporary science comes equipped with few options to pursue in order to accommodate a non-computational component. Only in one domain do non-computational processes play an irreducible role in the physical dynamics; that is, in quantum theory. The non-computational processes that punctuate the unitary evolution of quantum systems are variously termed 'wavefunction collapses', 'reductions of the wavepacket', 'measurements' or 'decoherence events', depending on the interpretation to which one subscribes, but their role in the dynamical descriptions of quantum theory is independent of interpretational issues and essential to the success of quantum descriptions of nature.

The suggestion that something as seemingly distant from our conception of mind and brain as quantum theory should enter into the explanation of consciousness has ignited a firestorm of controversy, much of the criticism proceeding on the *intuition* that an understanding of consciousness should not involve anything as esoteric as quantum theory. This kind of argumentation does a disservice to science by replacing sober reason and reflection with the weight of popular opinion. Similarly insinuations that Penrose is merely conflating the mystery of consciousness with the mysteries of quantum theory willfully ignore the logical basis of the argument and treat instead a caricature of the argument as unmotivated speculation.

Some of the reactions however pinpoint substantial issues and must be addressed within the context of any research project pursuing a putative link between quantum theory and a scientific understanding of the basis of consciousness. As one might expect, since the issue of quantum theory has been raised, some of the dissenting opinions originate from physics. These views target the disruptive effect of thermal noise and question the plausibility of instantiating a causally efficacious quantum system in the hot and noisy environment afforded by the living brain. Considerations along these lines must certainly condition any and all theories that attempt to provide realizable accounts involving a quantum element. They do not however present an impenetrable barrier. Quantum coherence is after all a commonplace at room temperature thanks to the rapid advancement of laser technology. In any case, it is not the current state of technology that should determine the course of scientific investigation but rather principled argument and a process of consultation with nature. It was this sort of process that guided the formulation of quantum theory in the first place, through the pioneering proposals of luminaries like Planck and de Broglie. *A priori* a proof concerning the mathematical *possibility* of a purely computational theory of consciousness would seem to take precedence over physical *plausibility* arguments based on the current limitations of our ability to manipulate quantum systems. One is led to wonder whether any statements of a more general nature might be gleaned from a consideration of consciousness from the perspective of physical science.

4 Physical Considerations

It seems clear that a conscious state, whatever its ultimate explanation, must incorporate numerous elements and establish specific relationships amongst them. It is also evident, on the basis of our current understanding of the brain, that there is no physical location on which these elements converge, that different elements composing a single conscious state are associated with the neurophysiological features of separated sites in the brain. The validity of these assertions is widely recognized and, taken together, they give rise to the *binding problem* which famously plagues the brain sciences: how do the neural activities in disparate physical locations come to be composed in a unified conscious state?

To illustrate the difficulty from a physical perspective and to discuss the alternative methods by which one might address the problem, it will be convenient to adopt a characterization of the brain due to Stapp[20]. One can imagine in principle modeling the brain to arbitrary precision by a vast network of devices, one for each of the nodes in a fine lattice of spacetime points covering the brain, each evolving through a series of arbitrarily small time steps.

The device at each node need only be sufficiently sophisticated to record the values of a handful of elementary matter and gauge fields and their first few derivatives, and to update those values on the basis of information passed at each time step from the nearest-neighbour nodes around it. Such a model is in keeping with the requirement that any physical system amenable to classical description can be decomposed as a collection of simple and independent entities, interacting only in accordance with a locality constraint.

On the basis of this model, one can contrast two distinct modes of description for the state of the system at a given time step. The first, an *intrinsic* mode, is merely the accumulation of all the values recorded in each of the devices. This is the mode of description appropriate to classical physics. It includes no information interpreting the collection of facts and no explicit representation of any relationships that might exist amongst the elements. In particular, functional units determined as aggregates of the elements play no role in such a description; these are part of an external interpretation, not fundamental entities in the clockwork model we have envisioned. They can have no causal effect on the physical processes occurring in the brain model and are irrelevant to its operation in the intrinsic mode.

Alternatively, the model might be described in a second mode, an *extrinsic* mode. This is the sort of description that might be formed by an external observer capable of *simultaneously* apprehending all of the values recorded at each of the nodes. This mode of description can be characterized by a *single* device capable of specifying the *entire* collection of recorded values, without the need to gather and compose the information from physically separated locations.

The complexity of a conscious state argues against the possibility that any one of the devices at the nodes of the lattice might alone be sufficient to encapsulate the requisite information. As the lattice tends towards a finer and finer covering of the volume, the information recorded by each node tends to something small in accordance with the precepts of classical physics. In any case, as noted above, there is little need to belabor the point as there is no support in the neurophysiological data currently available for the thesis that consciousness arises from a single location in the brain. The complexity of conscious states must therefore arise out of information from several, presumably many, nodes. But the conscious apprehension of the collection of facts occurs 'at once'. They are perceived as a single whole. The binding problem can thus be reduced to the problem of explaining an extrinsic level phenomenon in the intrinsic level language of classical physics. Consideration of this problem has led some to speculate concerning a principle of 'emergence'. Certainly *some* principle must supplement the edicts of classical physics but a scientific theory must go further, supplying not only a name but also attempting an explanation of the intuited features of consciousness.

The distinction between the intrinsic and extrinsic modes of description is constituted in the accessibility of relational data. Relations between elements that can be explicit in the extrinsic mode must remain implicit in the intrinsic mode. Simultaneous access to relational information is the crucial determinant of an extrinsic level description. The intuition in the computationalist camp is that the information to support consciousness can be accumulated in a piecemeal fashion, that it need not be made available simultaneously, that information can be stored throughout the brain and composed little by little. This intuition conceals the implicit *assumption* that the ordering relations inherent in spacetime can be used to provide the relational information necessary to reconstruct an extrinsic level description. The computationalist imagines 'loading' information into an abstract 'space'. On a classical account, the 'space' involved cannot be physical spacetime or we are reduced to an intrinsic mode of description. But if it is not actual spacetime then what supports the ordering relations and how are they maintained?

Within the scope of computational problems generally considered the assumption that physical spacetime should provide ordering relations has almost indiscriminate validity since computational problems do not generally require an extrinsic mode to be invoked. All forms of digital computation 'borrow' relational data from the structure of spacetime whenever 'memory' is invoked or whenever 'output' is written. This is independent of specific architectures. It applies, not only to standard von Neumann computers but also to neural networks and other parallel architectures and can be seen in the general context of a universal Turing machine. In a Turing machine the ordering of spacetime is concealed within the structure of the 'tape' and the 'borrowing' of the inherent relational information is disguised within the standard command, "Advance the tape." This instruction to the Turing machine brings a particular 0 or 1 under the reading device which particular 0 or 1 is determined by the structure of the tape itself. It is the tape that preserves the relation of every 0 and 1 written there to every other 0 and 1. It is this structure that underlies and supports the ordering of 0's and 1's on a Turing machine tape and likewise 'memory storage' and 'output' in a real computer.

These considerations introduce a dilemma. If conscious states are states in physical spacetime, then the imposition of a locality constraint eliminates the possibility of extrinsic level phenomena[1]. If conscious states are in some way

[1]Moreover, a *homonculus problem* is introduced. If the visual *perception* of a scene for instance is something that occurs physically in the brain, then we would seem to be requiring that there be a little man, a homonculus, in the head to perceive the perception. This cannot

'unphysical', if we take seriously the 'mental' features of psychological states, then the ordering of spacetime is not available to provide the relational information necessary to an extrinsic description and we require therefore some *other* space to perform the same function. The computationalist view of consciousness, a view which adopts the latter strategy, appears to be covertly dualist and would therefore seem to be beset by the same problem which plagues any dualist theory: how do the two separate spaces interact without violating conservation theorems? Apparently the only way to avoid the extravagant introduction of an entirely new space and its attendant difficulties is to relax the locality constraint. Dropping this requirement allows extrinsic level descriptions to be expressed within physical spacetime but also necessarily transports us from the classical realm to a quantum domain. Physical considerations then appear to point in the same direction as Penrose's mathematical result but arrive at the same conclusion by an independent route. An explanation of the specifically conscious aspects of what the brain does should require some quantum component.

5 Outlook

The arguments considered here do not indicate how a quantum element might be invested into a theory of consciousness but several avenues of research are currently being explored[6, 7, 8, 9, 10, 11, 12, 13, 14, 23, 24]. Recent work on quantum computation, for example, has uncovered an entirely new paradigm[3] that offers the intriguing possibility of biological instantiation. The pure quantum states required for quantum computation have generally been deemed much too delicate to survive amidst the thermal noise of biological environments. It has recently been demonstrated however that the small deviations from equilibrium observed in the quantum description of a macroscopic ensemble can be effectively treated as a *pure* quantum state of much lower dimension. Rather than isolating the quantum system from the environment to preserve its coherence, a pure state is *statistically* recreated from an unprotected mixed state. This work makes quantum computation an experimental reality at room temperature and presents a feasibility argument for quantum systems in biological media.

Similarly, models involving macroscopic ordered states have attempted to bridge the gulf that currently lies between abstract arguments for the indispensability of quantum theory in explaining consciousness and the concrete

constitute an explanation of consciousness then since it introduces a conscious entity in the explanation and an infinite regress thereby ensues.

realities of biology[1, 2, 21, 22]. These models escape the destructive influence of a thermal environment by maintaining a constant supply of energy, in much the same way that lasers maintain a state of quantum coherence at room temperature. The quantum system is located subcellularly in the network of microtubulin that supplies the cell with energy. Biological instantiations of this kind have begun to speculate about the manner in which living systems might take advantage of quantum variables to achieve an informational capacity much greater than classical counterparts.

That the explanation of the complex machinery of consciousness should find a subcellular component is independently suggested by the site of action of general anesthetics[5]. Medical researchers have accumulated a wide array of chemical substances that reversibly ablate consciousness. Physicochemical similarities amongst these substances implicate non-polar globular proteins as the site of action, specifically hydrophobic pockets in membrane lipids and microtubulin. Since the binding of general anesthetics does not change the membrane potential, it is hard to see, on a purely neural account, how they might alter the functioning of the brain so profoundly as to remove consciousness. This might find more natural explanation in a subcellular context.

Certainly our intuition rebels against the idea that something so seemingly foreign to our concept of the brain should play a necessary role in the generation of consciousness. And certainly our current understanding of the brain provides no convenient niche for the installation of a quantum connection. But just as certainly, it is through the constructive interplay of theoretical considerations and experimental findings that science is advanced.

References

[1] E. Del Guidice, S. Doglia, M. Milani and G. Vitiello, *Electromagnetic field and spontaneous symmetry breaking in biological matter*, Nucl. Phys. B275, 185 (1986).

[2] E. Del Guidice, S. Doglia, M. Milani and G. Vitiello, *A quantum field theoretical approach to the collective behavior of biological systems*, Nucl. Phys. B251, 375 (1985).

[3] N. A. Gerschenfeld and I. L. Chuang, *Bulk spin-resonance quantum computation*, Science 275, 350 (January 17, 1997).

[4] K. Gödel, *Über formal unentscheidbare Sätze per Principia Mathematica und verwandter Systeme I*, Monatshefte für Mathematik und Physik 38, 173 (1931).

[5] S. Hagan, M. Jibu and K. Yasue, *Consciousness and anesthesia: an hypothesis involving biophoton emission in the microtubular cytoskeleton of the brain* in *Origins: Brain and Self-organization*, K. H. Pribram (ed.), (Lawrence Erlbaum, New Jersey, 1994).

[6] S. Hameroff, R. Penrose, *Conscious events as orchestrated space-time selections*, J. Consciousness Studies 3, 36 (1996).

[7] S. Hameroff, R. Penrose, *Orchestrated reduction of quantum coherence in brain microtubules: a model for consciousness* in *Toward a Science of Consciousness - The First Tucson Discussions and Debates*, S. R. Hameroff, A. Kaszniak and A. Scott (eds.), (The MIT Press, Cambridge, Mass., 1996).

[8] S. R. Hameroff, *Quantum coherence in microtubules: A neural basis for emergent consciousness?*, J. Consciousness Studies, 1, 91 (1994).

[9] M. Jibu, K. H. Pribram and K. Yasue, *From conscious experience to memory storage and retrieval: the role of quantum brain dynamics and boson condensation of evanescent photons*, Int. J. Mod. Phys. B10, 1735 (1996).

[10] M. Jibu and K. Yasue, *Quantum Brain Dynamics - An Introduction*, (John Benjamin, Amsterdam, 1995).

[11] M. Jibu, S. Hagan, K. Pribram, S. R. Hameroff and K. Yasue, *Quantum optical coherence in cytoskeletal microtubules: implications for brain function*, BioSystems 32, 195-209 (1994).

[12] M. Jibu and K. Yasue, *Introduction to quantum brain dynamics* in *Nature, cognition ans system III*, E. Carvallo (ed.), (Kluwer Academic, London, 1993).

[13] M. Jibu and K. Yasue, *Intracellular quantum signal transfer in Umezawa's quantum brain dynamics*, Cybern. Syst. Int. J. 24, 1 (1993).

[14] M. Jibu and K. Yasue, *A physical picture of Umezawa's quantum brain dynamics* in *Cybernetics and Systems Research '92*, R. Trappl (ed.), (World Scientific, Singapore, 1992)).

[15] R. Penrose, *Shadows of the Mind*, (Oxford University Press, New York, N.Y., 1994).

[16] H. Putnam, *Brains and behavior* in *Analytical Philosophy*, R. Butler (ed.), (Basil Blackwell, Oxford, 1963).

[17] G. Ryle, *The Concept of Mind*, (Barnes and Noble, New York, 1949).

[18] J. Searle, *The Rediscovery of the Mind*, (The MIT Press, Cambridge, Mass., 1994).

[19] J. Searle, *Minds, brains, and programs*, Behavioral and Brain Sciences 3, 417 (1980).

[20] H. P. Stapp, *Why classical mechanics cannot naturally accommodate consciousness but quantum mechanics can*, Psyche (1995).

[21] C. I. J. M. Stuart, Y. Takahashi and H. Umezawa, *Mixed system brain dynamics: neural memory as a macroscopic ordered state*, Found. Phys. 9, 301 (1979).

[22] C. I. J. M. Stuart, Y. Takahashi and H. Umezawa, *On the stability and non-local properties of memory*, J. Theor. Biol. 71, 605 (1978).

[23] G. Vitiello, *Living matter physics and the quantum brain model*, Physics Essays 9, 538 (1996).

[24] G. Vitiello, *Dissipation and memory capacity in the quantum brain model*, Int. J. Mod. Phys. B9, 973 (1995).